Вероятность переобучения плотных и разреженных многомерных сеток алгоритмов*

Фрей А. И.
sashafrey@gmail.com
Московский Физико-технический институт

Известно, что для получения высокоточных оценок обобщающей способности в задачах обучения по прецедентам необходимо вводить более «тонкие» характеристики семейства алгоритмов, чем размерность. Предлагаются две такие характеристики — высота и разреженность. Показывается, что для монотонных и унимодальных многомерных сеток алгоритмов этих характеристик достаточно для получения точной оценки вероятности переобучения. Исследуется зависимость вероятности переобучения от размерности, высоты и разреженности для метода рандомизированной минимизации эмпирического риска.

The probability of overfitting for dense and sparse multidimensional grids of classifiers*

Frei A. I.
Moscow Institute of Physics and Technology, Moscow, Russia

The dimensional characteristics of the hypotheses set are known to be insufficient for obtaining tight generalization bounds. We propose two novel characteristics — the height and the sparsity of the hypotheses set, and show that they are sufficient to obtain exact generalization bounds for two special sets — the monotonic and unimodal multidimensional grids of classifiers. Then we study how the probability of overfitting depends on dimension, height, and sparsity in the case of randomized empirical risk minimization.

Введение

Для построения надёжных методов обучения по прецедентам необходимо иметь как можно более точные оценки обобщающей способности, например, в виде верхних оценок вероятности переобучения. Известные оценки используют различные характеристики семейства алгоритмов, метода обучения и обучающей выборки [4], в первую очередь, сложность (размерность) семейства алгоритмов. Точность таких оценок может быть недостаточной для практических применений [6]. Показано [7, 8], что надёжное обучение возможно только для семейств, обладающих одновременно двумя свойствами — расслоением алгоритмов по частотам ошибок и сходством алгоритмов.

Эффект расслоения состоит в том, что при фиксированной обучающей выборке малую частоту ошибок имеет лишь малая доля алгоритмов. Вероятность получить алгоритм в результате обучения резко падает с ростом его частоты ошибок. Поэтому достаточно оценивать сложность лишь нижних слоёв семейства, иначе оценки вероятности переобучения окажутся сильно завышенными.

Эффект сходства состоит в том, что схожие алгоритмы вносят существенно меньший вклад в вероятность переобучения, чем несхожие. Поэтому вероятность переобучения у семейств, непрерывных по параметрам, может оказаться вполне приемлемой даже при высокой размерности.

В [7, 8] было также показано, что учёт этих двух эффектов по отдельности в общем случае не даёт численно точных оценок вероятности переобучения. Но при этом не было предложено каких-либо удобных (желательно, скалярных) численных характеристик расслоения и сходства. Векторную характеристику *профиля расслоения* пока удавалось оценивать только с помощью весьма трудоёмкого метода Монте-Карло [2].

В данной работе вводятся скалярные характеристики *высоты* и *разреженности* семейства алгоритмов, связанные со свойствами расслоения и сходства, соответственно. Рассматриваются модельные семейства — многомерные монотонные и унимодальные сетки алгоритмов, для которых этих двух характеристик (в дополнение к длине выборке и размерности) оказывается достаточно для получения точной оценки вероятности переобучения. Эти семейства впервые были рассмотрены в [1], а их одномерные частные случаи — монотонные и унимодальные цепочки алгоритмов — в [7, 8]. Отличие данной работы в том, что вводится понятие разреженности семейства и применяется теоретико-групповой подход [3] для рандомизированного (а не пессимистичного, как в [1]) метода минимизации эмпирического риска.

Таким образом, показана принципиальная возможность получения точных оценок вероятности переобучения, в которых свойства размерности, расслоения и сходства выражаются тремя скалярными характеристиками, соответственно.

Работа поддержана РФФИ (проект № 08-07-00422) и программой ОМН РАН «Алгебраические и комбинаторные методы математической кибернетики и информационные системы нового поколения».

Постановка задачи

Пусть $\mathbb{X} = (x_i)_{i=1}^L$ — генеральная выборка, состоящая из L объектов. Отображения $a\colon \mathbb{X} \to \{0,1\}$ будем называть алгоритмами и говорить, что алгоритм a допускает ошибку на объекте x_i, если $a(x_i) = 1$. Каждому алгоритму взаимно однозначно соответствует бинарный вектор ошибок $\big(a(x_i)\big)_{i=1}^L$.

Величина $n(a,U) = \sum\limits_{x\in U} a(x)$ называется *числом ошибок* алгоритма a на подвыборке $U \subseteq \mathbb{X}$.

Величина $\nu(a,U) = n(a,U)/|U|$ называется *частотой ошибок* алгоритма a на подвыборке $U \subseteq \mathbb{X}$.

Обозначим через $\mathbb{A} = \{0,1\}^L$ множество всех 2^L бинарных векторов длины L, тогда $2^{\mathbb{A}}$ — это множество всех подмножеств $\mathbb{A}$.

Обозначим через $[\mathbb{X}]^\ell$ множество всех ℓ-элементных подмножеств генеральной выборки $\mathbb{X}$.

Детерминированным методом обучения назовем произвольное отображение $\mu\colon 2^{\mathbb{A}} \times [\mathbb{X}]^\ell \to \mathbb{A}$, которое по обучающей выборке $X \in [\mathbb{X}]^\ell$ выбирает из подмножества $A \subseteq \mathbb{A}$ некоторый алгоритм $a = \mu(A,X)$. Метод μ называется *минимизацией эмпирического риска* (МЭР), если выбираемый им алгоритм допускает наименьшее число ошибок на обучении: для всех $X \in [\mathbb{X}]^\ell$ и $A \subseteq \mathbb{A}$

$$\mu(A,X) \in A(X) \equiv \operatorname{Arg}\min_{a\in A} n(a,X).$$

Вопрос о том, какой именно алгоритм a из $A(X)$ выдать в результате обучения, может решаться по-разному. В пессимистичном методе МЭР выбирается алгоритм с максимальным $n(a,\mathbb{X})$, что приводит к верхним оценкам вероятности переобучения [8, 1]. Мы рассматриваем *рандомизированный метод обучения* [3, 5], который произвольным $A \in 2^{\mathbb{A}}$ и $X \in [\mathbb{X}]^\ell$ ставит в соответствие не один алгоритм a, а нормированную функцию $f(a)$, значение которой можно интерпретировать как вероятность получить алгоритм a в результате обучения:

$$\mu\colon 2^{\mathbb{A}} \times [\mathbb{X}]^\ell \to \Big\{ f\colon \mathbb{A} \to [0,1] \;\Big|\; \sum_{a\in\mathbb{A}} f(a) = 1 \Big\}. \quad (1)$$

Примером рандомизированного метода обучения является *рандомизированный метод МЭР*, основанный на равновероятном выборе a из $A(X)$:

$$\mu(A,X)(a) = \big[a \in A(X)\big] / \big|A(X)\big|. \quad (2)$$

Тут и далее квадратные скобки переводят логическое выражение в число: [истина] = 1, [ложь] = 0.

Пусть $X \sqcup \bar{X} = \mathbb{X}$ — произвольное разбиение генеральной выборки на обучающую выборку $X \in [\mathbb{X}]^\ell$ и контрольную $\bar{X} = \mathbb{X}\backslash X$. *Уклонением частот* назовем разность частот ошибок на контроле и на обучении: $\delta(a,X) = \nu(a,\bar{X}) - \nu(a,X)$.

Зафиксируем параметр $\varepsilon \in (0,1]$.

Будем говорить, что алгоритм a *переобучен* при разбиении $X \sqcup \bar{X}$, если $\delta(a,X) \geqslant \varepsilon$.

Примем единственное вероятностное предположение, что все разбиения генеральной выборки $\mathbb{X}$ на две подвыборки — наблюдаемую обучающую X и скрытую контрольную $\bar{X}$ — равновероятны [8]. Тогда *вероятность переобучения* для детерминированного метода обучения μ определяется как

$$Q_\mu(\varepsilon, A) = \frac{1}{C_L^\ell} \sum_{X\in[\mathbb{X}]^\ell} \big[\delta(\mu(A,X),X) \geqslant \varepsilon\big],$$

а для рандомизированного метода μ — как

$$Q_\mu(\varepsilon, A) = \frac{1}{C_L^\ell} \sum_{X\in[\mathbb{X}]^\ell} \sum_{a\in A} \mu(A,X)(a) \big[\delta(a,X) \geqslant \varepsilon\big].$$

Монотонная сетка алгоритмов

Пусть $\boldsymbol{d} = (d_1,\ldots,d_h) \in \mathbb{Z}^h$ — целочисленный вектор индексов. Обозначим $\|\boldsymbol{d}\| = \max\limits_{j=1,\ldots,h} |d_j|$, $|\boldsymbol{d}| = |d_1| + \ldots + |d_h|$. На множестве векторов индексов введём покомпонентное отношение сравнения: $\boldsymbol{d} \leqslant \boldsymbol{d}'$, если $d_j \leqslant d'_j$, $j = 1,\ldots,h$; и $\boldsymbol{d} < \boldsymbol{d}'$ если хотя бы одно из неравенств $d_j \leqslant d'_j$ строгое.

Определение 1. *Множество алгоритмов* $A_M = \big\{a_{\boldsymbol{d}}\colon \boldsymbol{d} \geqslant \boldsymbol{0},\ \|\boldsymbol{d}\| \leqslant D\big\}$ *называется монотонной* h*-мерной сеткой алгоритмов высоты* D*, если* $\mathbb{X}$ *разбивается на непересекающиеся подмножества* U_0, U_1 *и* $X_j = \{x_j^1,\ldots,x_j^D\}$, $j = 1,\ldots,h$, *такие, что:*

1) $a_{\boldsymbol{d}}(x_j^i) = [i \leqslant d_j]$, *где* $x_j^i \in X_j$;
2) $a_{\boldsymbol{d}}(x_0) = 0$ *при всех* $x_0 \in U_0$;
3) $a_{\boldsymbol{d}}(x_1) = 1$ *при всех* $x_1 \in U_1$.

Монотонная сетка — это модель семейства алгоритмов с h непрерывными параметрами, предполагающая, что по мере непрерывного увеличения j-го параметра при фиксированных остальных ошибки возникают последовательно на объектах $x_j^1,\ldots,x_j^D$.

Обозначим $m \equiv |U_1|$. Из определения 1 следует, что $n(a_{\boldsymbol{d}},\mathbb{X}) = m + |\boldsymbol{d}|$. Число алгоритмов в h-мерной монотонной сетке высоты D составляет $(D+1)^h$. Алгоритм $a_{\boldsymbol{0}}$ является *лучшим в сетке*.

Пример 1. Двумерная ($h = 2$) монотонная сетка при $m = 0$ и $L = 4$:

	$a_{0,0}$	$a_{1,0}$	$a_{2,0}$	$a_{0,1}$	$a_{1,1}$	$a_{2,1}$	$a_{0,2}$	$a_{1,2}$	$a_{2,2}$
x_1	0	**1**	**1**	0	**1**	**1**	0	**1**	**1**
x_2	0	0	**1**	0	0	**1**	0	0	**1**
x_3	0	0	0	**1**	**1**	**1**	**1**	**1**	**1**
x_4	0	0	0	0	0	0	**1**	**1**	**1**

Определение 2. *Разреженностью* ρ *множества алгоритмов* A *назовём минимальное хэммингово расстояние между векторами ошибок:*

$$\rho = \min_{a,a'\in A} \sum_{i=1}^L \big|a(x_i) - a'(x_i)\big|.$$

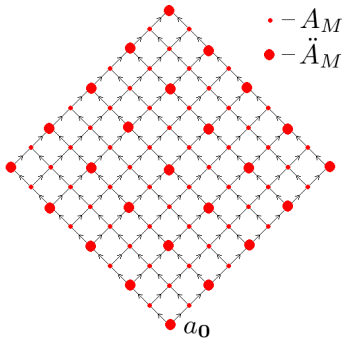

Рис. 1. Разреженная монотонная сетка. Узлы сетки соответствуют алгоритмам, направление стрелок — возрастанию числа ошибок алгоритмов.

Определение 3. *Разреженной h-мерной монотонной сеткой с разреженностью ρ называется подмножество* $\ddot{A}_M = \{a_{\boldsymbol{d}} \in A_M : \boldsymbol{d} \in (\rho\mathbb{Z})^h\}$.

Если исходная сетка A_M имела высоту D, то величину $\lfloor D/\rho \rfloor$ будем называть высотой сетки $\ddot{A}_M$.

Пример 2. На рис. 1 выделено подмножество алгоритмов двумерной монотонной сетки высоты $D = 8$, соответствующее разреженной монотонной сетке с параметрами $\rho = 2$, $D = 4$.

Введем дополнительные обозначения:

C_n^k — биномиальные коэффициенты, причём будем считать, что $C_n^k = 0$ при $k < 0$ и $k > n$;

$H_L^{\ell,m}(z) = \frac{1}{C_L^\ell} \sum\limits_{s=0}^{\lfloor z \rfloor} C_m^s C_{L-m}^{\ell-s}$ — функция гипергеометрического распределения, $z \in [0, \ell]$;

$Y_*^{h,D} \subset \mathbb{Z}^h$ — множество целочисленных невозрастающих неотрицательных последовательностей длины h, первый член которых не превосходит D;

S_h — группа перестановок элементов последовательности $\boldsymbol{\lambda} \in Y_*^{h,D}$;

$|S_h\boldsymbol{\lambda}|$ — мощность орбиты действия S_h на $\boldsymbol{\lambda}$.

Теорема 1. *Пусть $\ddot{A}_M$ — разреженная h-мерная монотонная сетка высоты D и разреженности ρ. Тогда вероятность переобучения $Q_\mu(\varepsilon, \ddot{A}_M)$ для рандомизированного метода минимизации эмпирического риска дается формулой*

$$Q_\mu(\varepsilon, \ddot{A}_M) = \sum_{\boldsymbol{\lambda} \in Y_*^{h,D}} \sum_{\substack{\boldsymbol{t} \geqslant \rho\boldsymbol{\lambda}, \\ \|\boldsymbol{t}\| \leqslant \rho D}} \frac{|S_h\boldsymbol{\lambda}|}{T(\lfloor \boldsymbol{t}/\rho \rfloor)} \frac{C_{L'}^{\ell'}}{C_L^\ell} H_{L'}^{\ell',m}\left(s_0(\boldsymbol{\lambda})\right),$$

где $\ell' = \ell - \sum\limits_{j=1}^{h} [t_j \neq \rho D]$, $k' = k - |\boldsymbol{t}|$, $L' = \ell' + k'$,
$T(\boldsymbol{t}) = \prod\limits_j (t_j + 1)$, $s_0(\boldsymbol{\lambda}) = \frac{\ell}{L}\left(m + \rho|\boldsymbol{\lambda}| - \varepsilon k\right)$.

При $\rho = 1$ данная формула дает вероятность переобучения не-разреженной сетки алгоритмов A_M.

Унимодальная сетка алгоритмов

Унимодальная сетка является более реалистичной моделью семейства с h непрерывными параметрами, по сравнению с монотонной сеткой. Предполагается, что непрерывное отклонение j-го параметра не только в бóльшую, но и в меньшую, сторону от оптимального значения приводит к увеличению числа ошибок.

Определение 4. *Множество алгоритмов $A_U = \{a_{\boldsymbol{d}} : \|\boldsymbol{d}\| \leqslant D\}$ называется унимодальной h-мерной сеткой алгоритмов высоты D, если $\mathbb{X}$ разбивается на непересекающиеся подмножества U_1, U_0, $X_j = \{x_j^1, \ldots, x_j^D\}$, $Y_j = \{y_j^1, \ldots, y_j^D\}$, $j = 1, \ldots, h$, такие, что:*

1) $a_{\boldsymbol{d}}(x_j^i) = [d_j > 0]\,[i \leqslant |d_j|]$, *где* $x_j^i \in X_j$;
2) $a_{\boldsymbol{d}}(y_j^i) = [d_j < 0]\,[i \leqslant |d_j|]$, *где* $y_j^i \in Y_j$;
3) $a_{\boldsymbol{d}}(x_0) = 0$ *при всех* $x_0 \in U_0$;
4) $a_{\boldsymbol{d}}(x_1) = 1$ *при всех* $x_1 \in U_1$.

Данное определение отличается от определения монотонной сетки отсутствием ограничения $\boldsymbol{d} \geqslant \mathbf{0}$. Число алгоритмов в h-мерной унимодальной сетке высоты D равно $(2D+1)^h$. Как и для монотонной сетки, число ошибок $n(a_{\boldsymbol{d}}, \mathbb{X})$ алгоритма $a_{\boldsymbol{d}}$ равно $m + |d|$, где $m \equiv |U_1|$.

Определение 5. *Разреженной h-мерной унимодальной сеткой с разреженностью ρ называется подмножество* $\ddot{A}_U = \{a_{\boldsymbol{d}} \in A_U : \boldsymbol{d} \in (\rho\mathbb{Z})^h\}$.

Если исходная сетка A_U имела высоту D, то величину $\lfloor D/\rho \rfloor$ будем называть высотой сетки $\ddot{A}_U$.

Обозначим через $n(\boldsymbol{\lambda})$ число ненулевых компонент последовательности $\boldsymbol{\lambda} \in Y_*^{h,D}$.

Теорема 2. *Пусть $\ddot{A}_U$ — разреженная h-мерная унимодальная сетка высоты D и разреженности ρ. Тогда вероятность переобучения $Q_\mu(\varepsilon, \ddot{A}_U)$ для рандомизированного метода минимизации эмпирического риска дается формулой*

$$Q_\mu(\varepsilon, \ddot{A}_U) = \sum_{\boldsymbol{\lambda} \in Y_*^{h,D}} \sum_{\substack{\boldsymbol{t} \geqslant \rho\boldsymbol{\lambda}, \\ \|\boldsymbol{t}\| \leqslant \rho D}} \sum_{\substack{\boldsymbol{t}' \geqslant \mathbf{0}, \\ \|\boldsymbol{t}'\| \leqslant \rho D}} \mathbb{S}(\boldsymbol{\lambda}, \boldsymbol{t}, \boldsymbol{t}'),$$

$$\mathbb{S}(\boldsymbol{\lambda}, \boldsymbol{t}, \boldsymbol{t}') = \frac{|S_h\boldsymbol{\lambda}| \cdot 2^{n(\boldsymbol{\lambda})}}{T(\lfloor \boldsymbol{t}/\rho \rfloor + \lfloor \boldsymbol{t}'/\rho \rfloor)} \frac{C_{L'}^{\ell'}}{C_L^\ell} H_{L'}^{\ell',m}\left(s_0(\boldsymbol{\lambda})\right),$$

$\ell' = \ell - \sum\limits_{j=1}^{h} \left([t_j \neq \rho D] + [t_j' \neq \rho D]\right)$, $k' = k - |\boldsymbol{t}| - |\boldsymbol{t}'|$,
остальные обозначения те же, что в теореме 1.

Вычислительный эксперимент

На рис. 2 показана зависимость вероятности переобучения h-мерной монотонной сетки от ее высоты. При увеличении высоты сетки свыше $D = 5$ вероятность переобучения выходит на константу. Поэтому дальнейшие графики построены при малом значении параметра $D = 3$.

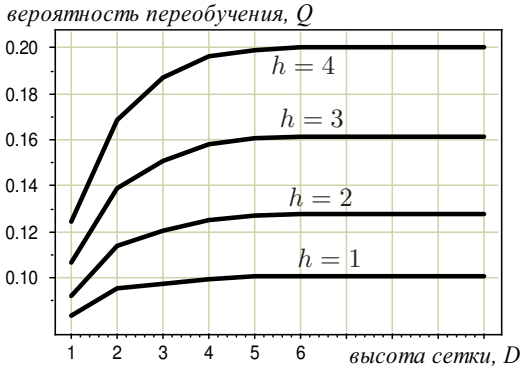

Рис. 2. Зависимость $Q_\mu(\varepsilon, \ddot{A}_M)$ от высоты D монотонной сетки при $L = 150,\ \ell = 90,\ \varepsilon = 0.05,\ m = 5,\ \rho = 1$.

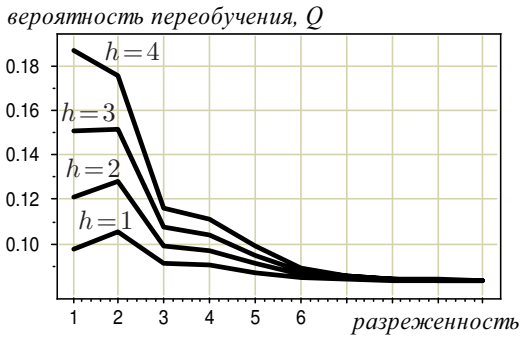

Рис. 3. Зависимость $Q_\mu(\varepsilon, \ddot{A}_M)$ от разреженности ρ при $L = 150,\ \ell = 90,\ \varepsilon = 0{,}05,\ D = 3,\ m = 5$.

На рис. 3 изображена зависимость вероятности переобучения h-мерной монотонной сетки при $h = 1, 2, 3, 4$ от разреженности ρ. При увеличении разреженности ρ вероятность переобучения падает, и вскоре выходит на константу, соответствующую вероятности переобучения лучшего алгоритма семейства a_0. Это связано с тем, что с увеличением ρ вероятность получить в результате обучения алгоритм a_0 стремится к единице.

На рис. 4 приведены результаты сравнения разреженных h-мерных унимодальных сеток с разреженными $2h$-мерными монотонными сетками, при $h = 1$ и $h = 2$. Серая кривая соответствует вероятности переобучения унимодальной сетки. Результаты подтверждают гипотезу [1] о том, что вероятность переобучения унимодальной сетки и монотонной сетки двойной размерности очень близки.

Выводы

В работе предложены два новых параметра множества алгоритмов — *разреженность* и *высота*. Для разреженных монотонных и унимодальных сеток показано, что эти параметры, наряду с длиной выборки и размерностью, определяют вероятность переобучения. Получены точные формулы вероятности переобучения для плотных ($\rho = 1$) и разреженных ($\rho > 1$) многомерных монотонных и унимодальных сеток в случае рандомизированной минимизации эмпирического риска. С помощью полученных формул экспериментально установлено, что в широком диапазоне параметров для монотонных и унимодальных сеток вероятность переобучения определяется несколькими нижними слоями семейства. Также показано, что с увеличением разреженности семейства вероятность получить лучший алгоритм в результате обучения стремится к единице.

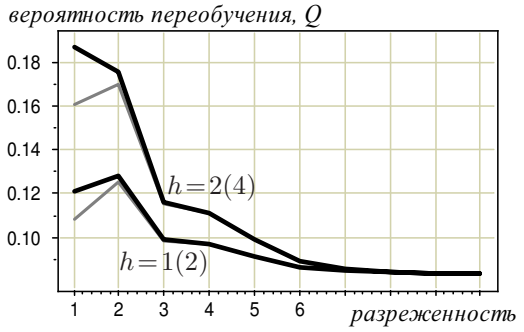

Рис. 4. Сравнение $Q_\mu(\varepsilon, \ddot{A}_M)$ и $Q_\mu(\varepsilon, \ddot{A}_U)$ от ρ при $L = 150,\ \ell = 90,\ \varepsilon = 0{,}05,\ D = 3,\ m = 5,\ h = 1(2), 2(4)$.

Литература

[1] *Ботов П. В.* Точные оценки вероятности переобучения для монотонных и унимодальных семейств алгоритмов // Всеросс. конф. ММРО-14. — М.: МАКС Пресс, 2009. — С. 7–10.

[2] *Кочедыков Д. А.* Структуры сходства в семействах алгоритмов классификации и оценки обобщающей способности // Всеросс. конф. ММРО-14. — М.: МАКС Пресс, 2009. — С. 45–48.

[3] *Фрей А. И.* Точные оценки вероятности переобучения для симметричных семейств алгоритмов // Всеросс. конф. ММРО-14. — М.: МАКС Пресс, 2009. — С. 66–69.

[4] *Boucheron S., Bousquet O., Lugosi G.* Theory of classification: A survey of some recent advances // ESAIM: Probability and Statistics. — 2005. — no. 9. — Pp. 323–375.

[5] *Frey A. I.* Accurate Estimates of the Generalization Ability for Symmetric Sets of Predictors and Randomized Learning Algorithms // Pattern Recognition and Image Analysis. — 2010. — Vol. 20, no. 3. — P. 241–250.

[6] *Vorontsov K. V.* Combinatorial probability and the tightness of generalization bounds // Pattern Recognition and Image Analysis. — 2008. — Vol. 18, no. 2. — Pp. 243–259.

[7] *Vorontsov K. V.* Splitting and similarity phenomena in the sets of classifiers and their effect on the probability of overfitting // Pattern Recognition and Image Analysis. — 2009. — Vol. 19, no. 3. — Pp. 412–420.

[8] *Vorontsov K. V.* Exact combinatorial bounds on the probability of overfitting for empirical risk minimization // Pattern Recognition and Image Analysis. — 2010. — Vol. 20, no. 3. — P. 269–285.